(2)

Redesigned Isis and System under Mach

Second Quarterly R & D Status Report
Apr. 1, 1993

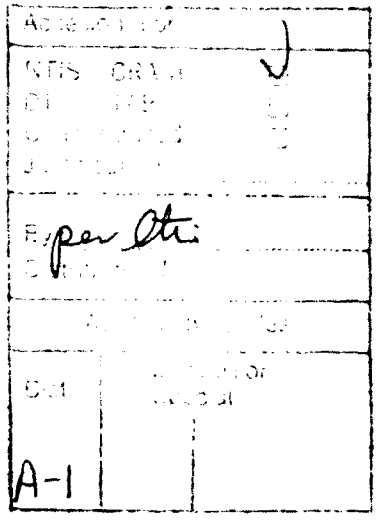
Prof. Kenneth P. Birman
Department of Computer Science
Cornell University, Ithaca New York
607-255-9199

Prof. Keith Marzullo
Dept. of Computer Science
U.C. San Diego, San Diego, California

This work was sponsored by the Defense Advanced Research Projects Agency (DoD), under contract N00014-92-J-1866 issued by the Office of Naval Research. The view, opinions and findings contained in this report are those of the authors and should not be construed as an official DoD position, policy, or decision.

Personnel

- Academic Staff:
  - Prof. Kenneth P. Birman, Principle-Investigator
  - Prof. Keith Marzullo, Co-Investigator
  - Dr. Robert Cooper, Research Associate
  - Dr. Robbert van Renesse, Research Associate
  - Dr. Aleta Ricciardi, Post-doctoral Research Associate
  - Dr. Mark Wood, Post-doctoral Research Associate
- Graduate Students:
  - Lorenzo Alvisi (Marzullo)
  - Navin Budhiraja (Marzullo)
  - Brad Glade (Birman)
  - Guerney Hunt (Birman)
  - Neil Jain (Birman)
  - David Karr (Birman)
  - Michael Kalantar (Birman)
  - Michael Reiter (Birman)
  - Aleta Ricciardi (Birman)
  - Laura Sabel (Marzullo)



The Isis project

This quarterly status report covers activities of the Isis project during the second quarter of 1993. This is our second progress report under ONR funding, but because these status reports are intended to be brief and our proposal was recently funded, we assume that the reader has some background regarding the goals and status of our effort, and focus instead on technical accomplishments during the report period and goals for the next three months. Readers unfamiliar with our work could start by reading some of the papers cited below, such as TR 1216, which will appear in Communications of the ACM later this year.

During the report period, the Isis effort has achieved steady progress in its effort to redesign and reimplement the Isis system using Mach and Chorus as target operating system environments. This work was documented in two new technical reports, which have been submitted to the ACM SIGOPS Symposium on Operating Systems Principles. In addition, we integrated a key component of our new system into Mach, resulting in a substantial improvement in the fault-tolerance of Mach. Finally, our work on Meta and Corto resulted in a number of other publications that are described and listed below.

During 1993 our focus remains the completion of our new system, its full integration with Mach and Chorus, and other microkernel operating systems. Through joint work with UCSD and with INESC, Portugal, we also expect that a major effort in the real-time area will take shape during this period. We also plan to press for greater industry standardization around our approach, notably through a proposal we have made to the OMG organization.

The major accomplishments of this quarterly report period are as follows:

- We completed the first version of the Horus system, and have begun building a toolkit over it, similar to the one used on the older ISIS system. Predictions of a 10- to 100-fold performance improvement appear to be justified, but until we have this new software running under native Mach it will be difficult to say anything final on the issue. A port to native Mach is underway at this time.

- We integrated an important part of the new system (its failure agreement module) into the Mach system, giving Mach a way to recover from node failure safely, and without waiting for concrete proof that

the node was faulty. This represents a major improvement on the Mach side, and also a first step towards full integration of our system with Mach.

- We continued work on a new way of presenting Isis groups that will reduce costs by allowing Isis to map multiple application-level process groups to a single Isis process group. The idea here is to amortize membership changes over multiple groups so as to reduce their effective cost. The technique us expected to avoid high overhead in applications that use very large numbers of nearly identical groups.

- Continuation of research ties with other laboratories, including the Los Alamos Advanced Computing Laboratory (which focuses on supercomputing), Portugal's INESC research laboratory (known for its work on realtime communication), and with Mach-related research efforts at the Open Software Foundation, Carnegie Mellon, and University of Arizona. The goal in our work with LANL is currently to port Isis to the Cray YMP processor series. With INESC, we have defined a common system architectural layering and are now beginning to unify older Delta-4 technology with Horus.

- We completed the development and initial implementation of the new security architecture for Horus, which focuses on securing islands of Horus users within hostile networks, and on securing Horus abstractions even within these islands. We view this as an extremely important advance, because the previous system (Isis) was almost completely trusting of its users. The secure Horus architecture, in contrast, can tolerate arbitrary failures outside a collection of physically secure nodes, and supports a highly sophisticated trust, encryption and delegation architecture within an island of secured nodes. Implementation of this architecture is proving to be a cornerstone of our new system, and we are close to being able to support users.

uring the last three months, Keith Marzullo has been busy resettling down at UC San Diego. He has been setting up a laboratory environment for building Corto and for continuing work on Meta and Lomita. We are currently negotiating a price for a set of computers, which will be purchased using startup funds supplied by UC San Diego. The lab will be operational by mid-May.

The hardware and software configuration for the laboratory machines

was chosen by consulting with Doug Locke of IBM FSC and Lui Sha's group at the Software Engineering Institute. Both have advocated starting off using LynxOS as the initial operating system platform because it is POSIX compliant with respect to real-time scheduling and it supports a Unix-like programming environment. Furthermore, Sha's group has built a laboratory using similar equipment and are currently working on scheduling problems for fault-tolerant distributed computing. By having a similar platform, Sha and Marzullo hope to be able to share results of their research.

Marzullo plans to start porting Horus to UC San Diego as soon as the lab is on the air. The first pieces of Corto—clock synchronization, group membership and real-time atomic multicast—will be built this Summer. There are protocols in the literature for these pieces, and so we will choose known efficient protocols and build them on top of MUTS. This will then provide a platform with which we can start experimenting in order to determine the constraints that the "active replication" approach we are proposing will impose.

- We have made substantial progress in a new experimental effort to understand flow control problems on hardware multicast technologies such as ethernet, FDDI and token ring, and are extending our work to include next-generation technologies such as ATM. The goal of this effort is to develop effective flow-control algorithms for use within the Horus multicast protocols. So far, we have focused on collecting data concerning the behavior of the raw devices themselves, and have obtained fascinating and non-intuitive results concerning packet loss rates in a number of settings. These show that the most significant loss rates are for small packets sent in many-one or many-many situations. Low or zero loss occurs with large packets and for one-many patterns. This information will be used to develop algorithms that narrow in on the situations in which loss rates are highest, while remaining uninvolved in other situations. Such flow control algorithms are the key element limiting Horus performance on many systems, and development of this new flow control software will be a small but critical activity for us during the coming year.

- We have initiated a new project to explore specialized implementations of both Isis and Horus for the CM/5 and Intel Touchstone multiprocessors. This work is motivated by the impressive results of Berke-

ley's Split/C and Active Messages research, demonstrating that asynchronous communication can lead to tremendous performance gains on the most important emerging parallel processors. In a very exciting development, the main developer of the Active Messages system (Thorsten von Eiken) is expected to join our project next fall, as a faculty member in the Cornell Dept. of Computer Science. As we move Horus onto highly parallel platforms, we want to build our protocols in ways that exploit the hardware fully and minimize unnecessary work in software – work needed on networks but not on closely coupled machines. We are very pleased with this new direction.

- We proposed to the Object Management Group (a standards organization) that ISIS group mechanisms be adopted as part of a reliability architecture for the Common Object Request Broker Archiecture (CORBA). A copy of the proposal is attached.

- Finally, and last only because the effort is one that started recently, we have begun to explore the integration of realtime support into Isis, through a project called CORTO. Our goals are fairly modest for this effort, at least initially, because we wish to build something usable which we can later extend with sophisticated schedulers and other adjuncts. In the near term, CORTO will focus on adding periodic process groups and realtime group communication to Isis.

Second Budget Statement

a.	ARPA Order Number:	7019
b.	Contract Number:	N00014-92-J-1866
c.	Agent:	ONR
d.	Contract Title:	A Redesigned ISIS and Meta System Under Mach
e.	Organization:	Cornell University
f.	PIs:	Kenneth P. Birman and Keith Marzullo
g.	Actual Start Date:	9/30/92
h.	Expected End Date:	12/30/95
i.	Expected End Date if Options Exercised:	NA
j.	Total Price:	$3,137,518
k.	Spending Authority Provided So Far:	$1,281,331
l.	Expenditures through 3/93	$278,000
m.	Date When These Funds Will Be Fully Expended:	12/31/92
n.	Additional Funds Expected Per Contract (by FY):	FY94 $928,050 FY95 $928,137

Publications

Below, we reproduce a list of recent publications by the effort. A good general review of the project is TR 1216, soon to be published by the Communications of the ACM. We are also well advanced on a book that will collect the most important papers into a single volume.

- Understanding Partitions and the 'No Partition' Assumption. André Schiper, Aleta Ricciardi, Kenneth Birman. To appear 4th FT-DCS (Future Trends in Distributed Computing Systems), September 22-24, 1993, Lisbon, Portugal.

- Virtually-Synchronous Communication Based on a Weak Failure Suspector. André Schiper, Aleta Ricciardi, Kenneth Birman. To appear 23rd FTCS (Fault-Tolerant Computing Symposium), June 22-24, 1993, Toulouse, France.

- Fault-tolerant Programming using Process Groups. Robbert van Renesse and Ken Birman. 1993. To appear in *IEEE Distributed Open Systems in Perspective.*

- Fault-Tolerant Key Distribution. Mike Reiter, Ken Birman and Robbert van Renesse. January 1993. Available as Cornell Technical Report 93-1325; Revised version submitted to the Fourteenth ACM Symposium on Operating Systems Principles.

- Monitoring and Controlling Distributed Applications using Lomita. Keith Marzullo and Ida Szafranska. IEEE First International Workshop on Systems Management, 14-16. April 1993.

- FLIP: An Internetwork Protocol for Supporting Distributed Systems. M. F. Kaashoek, R. van Renesse, H. van Staveren, and A. S. Tanenbaum. *ACM Transactions on Computer Systems*, volume 11, number 1.

- Nonblocking and Orphan-Free Message Logging Protocols. Lorenzo Alvisi, Bruce Hoppe and Keith Marzullo. To appear 23rd FTCS (Fault-Tolerant Computing Symposium), June 22-24, 1993, Toulouse, France.

- Ozalp Babaoglu, Keith Marzullo and Fred B. Schneider. Priority Inversion and its Prevention. Accepted for publication in *Journal of Real-Time Systems*, volume 5, number 4.
- The Process Group Approach to Reliable Distributed Computing. Ken Birman. July 1991. Available as Cornell TR 91-1216; to appear in the *Communications of the ACM.*
- Light-Weight Process Groups. Bradford Glade, Ken Birman and Robert Cooper. *Proceedings of the OpenForum '92 Technical Conference*, 323-336. November 1992.